Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Looking at the categorical variables, I believe that the feature columns "season", "yr", and "holiday". Seasonality ("season") affects the count of bicycle rental, as colder weather and more dangerous condition road conditions dissuade people from renting a bicycle, and rental count during Spring (labelled 1) would be much lower.

Year ("yr") affects the count of bicycle rental, as the company was newly established in the US market in 2018 and it was only in 2019 when the company became more well-known, so the rental count in 2019 was much higher.

Lastly, holiday ("holiday") affects the count of bicycle rental, as bicycle rental would be much lower when it was not a holiday, as people would be staying at home more often.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

It is important to use parameter of drop_first = True if want to drop the first column of the dummy variable which is redundant. For instance, we want to create dummy variable if the house is furnished or not, and we will recode furnished as 1, and not furnished as 0. However, pd.get_dummies() will get two feature columns that are directly opposite of each other; thus, still correlated, and the first variable is unnecessary.

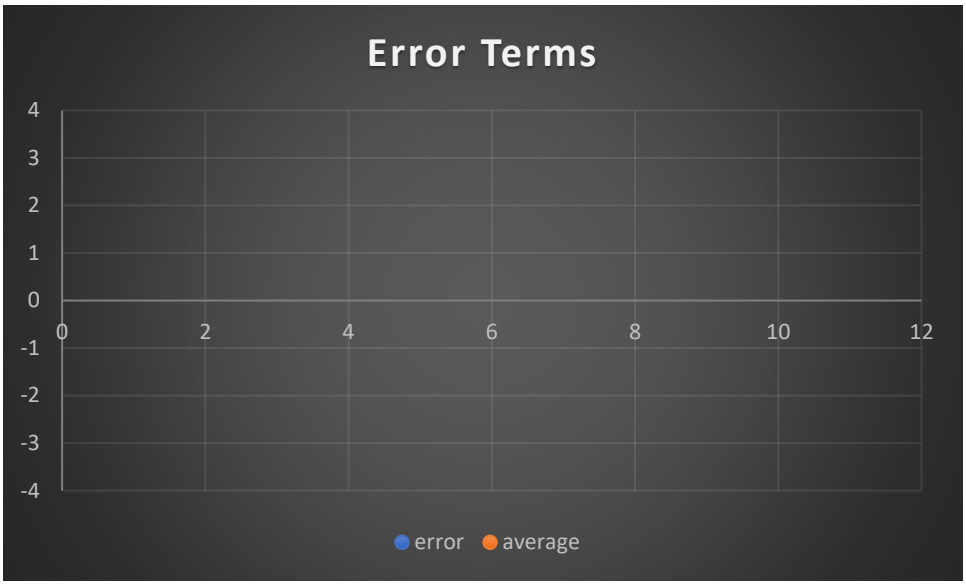
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

In observing the pair-plot, featured-names "registered" and "casual" both have the highest correlation with target variable. Reason being there is a data leak issue, and both feature columns are directly derived from the "cnt" column.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

The assumption of Linear Regression was validated by plotting a histogram to show the distribution of the error term. One of the assumptions of Linear Regression is that the error terms are normally distributed; therefore, if my error terms follow the shape of a normal distribution, I can confidently say that my model did not violate the assumption of Linear Regression.

Another way of validating the assumption of Linear Regression is by ensuring homoscedasticity of error terms. What this means it that the residuals should have constant variance, and the graph should look something like below ideally.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The top three contributing feature columns as below:

1. <u>Year.</u> This means that year is a important feature because more bikes are rented out during the second year (2019) when the company became more established and well-known in USA.

2. <u>Not Spring.</u> When the season is not spring, more bikes are rented out as Spring is a season when temperatures are lower and snow and dangerous weather conditions are more frequent. Note that this feature was featured engineered and not from the original dataset.

3. <u>Weather is not snow</u>. When the weather is snowing, far fewer bikes are rented out as it is safer to ride, and there is a strong inverse relationship, as indicated by high negative beta coefficient value. Note that this feature was featured engineered and not from the original dataset.

<u>General Subjective Questions</u>

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a machine learning algorithm in which we are interested to find out the effect of a variable has had on the target variable, how a change in the value of a variable will have on the impact of the target variable, and how we can explain the trend or the relationship between a dependent variable and an independent variable.

A linear regression, in this case a simple linear regression, is usually written in the formula y = mx + c or $Y = Beta_0 + Beta_1 * X$. A linear regression can be extended through a multiple linear regression, in which there will be more than one independent variable and can be written in the formula of $Y = Beta_0 + Beta_1 * X + Beta_2 * X + \ldots Beta_n * X$, in which there are up to n number of variables.

The strength of a simple linear regression model can be checked if the residual sum of squares (RSS) is extremely close to 0. The closer it is to 0, the better the fit the simple linear regression model is.

However, the strength of a multiple linear regression model is best checked through the adjusted R2 value, which is defined by the formula below:

R2 = 1-RSS/Total Sum of Squares (TSS).

Adjusted R2 = 1 – (1-R2)(N-1)/(N-p-1).

Where N refers to Total Sample Size, p refers to the number of independent variables.

Reason being R2 will be closer to 1 as we add in more independent variables, and adjusted R2 will be closer to 1 if the new independent variable added improves the model more than what would be by chance.

2. Explain the Anscombe's quartet in detail. (3 marks)

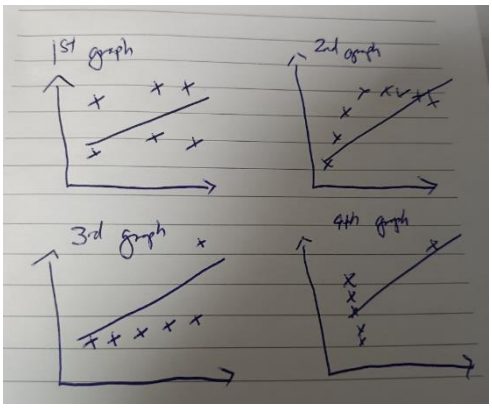
Anscombe's quartet refers to four datasets that have the same statistical properties (having the same mean for X-axis variables, standard deviation for X-axis variables, mean for Y-axis variables, standard deviation for Y-axis variables, and Pearson's R between X and Y), but it highlights the importance of visualizing the graph before making decisions.

In Anscombe's quartet, the first graph is when a straight line goes through the data points on the scatter plot and there is a linear relationship between X and Y.

The second graph is when X and Y have a non-linear relationship and the straight line cannot explain well the relationship between X and Y.

The third graph is when X and Y seemingly have a linear relationship but there is a point which is an outlier and appears far away from the best fit line.

The fourth graph is when X and Y have no linear relationship in reality, but there is an outlier and the best fit line fits the outlier into the graph and it becomes a straight line with high correlation coefficient is drawn.



3. What is Pearson's R? (3 marks)

The Pearson's R is a measure of the strength of relationship in a linear regression model between two variables. It is denoted by *r*. A best fit line is drawn between two variables and *r* measures the distance between the data points to the best fit line. R takes a value from -1 to 1, in which -1 refers to relationship between the two variables as in inversely perfectly correlated (one variable increases by

value of 1, the other variable decreases by value of -1), 0 refers to the variables having no relationship, and 1 refers to the variables being perfectly correlated (one variable increases by value of 1, the other variable increases by 1).

Pearson's R is usually used to measure the strength of relationship between two variables, unlike regression models in which the independent variable has an impact on the independent variable. For instance, for football players in the English Premier League, a player's transfer market value is usually correlated strongly with the players' weekly wages, the higher the weekly wages, the higher the transfer market value, because the transfer market value is usually the value that another club will pay for to "buy out" the contract of the player with his or her current club.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling refers to converting independent features and standardizing the values so that the values can be fixed within a fixed range.

Scaling is performed because of the various reasons: a) machine learning algorithm tends to consider larger values higher while at the same time, consider smaller values lower, b) scaling can help us to interpret the data easier as the data will be scaled down to a standardized, fixed range, and c) standardizing values can help us in converging faster during gradient descent.

The difference between normalized scaling and standardized scaling is in the values – normalizes scaling will not have a negative value, whereas standardized scaling will contain both positive and negative values as mean is at 0.

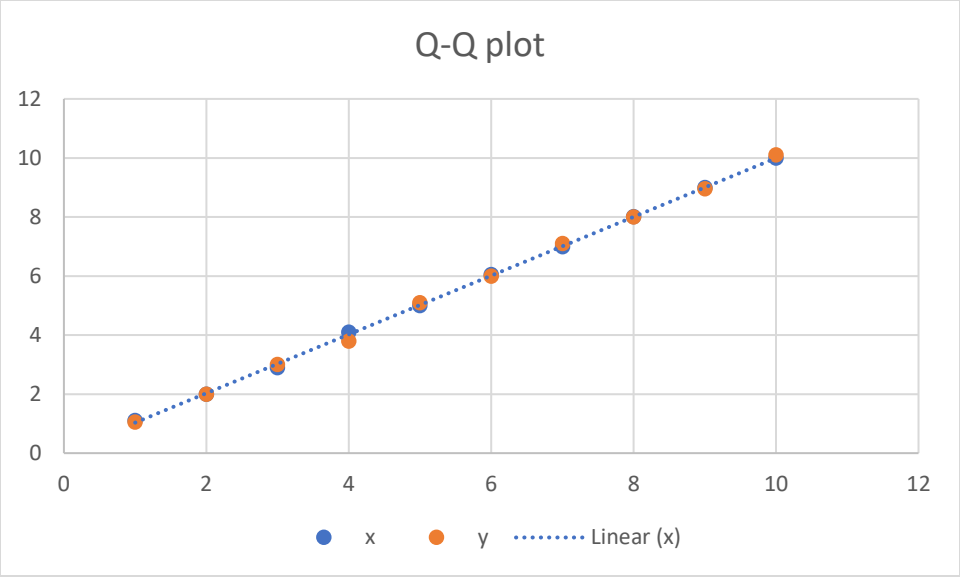
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Value of VIF can become infinite when there is perfect correlation between the two variables. This usually happens when the two variables are in fact the same. For example, we have one feature that is distance travelled in km, while we have another feature that is distance travelled in miles. This usually happens when the person who made the dataset might have decided to recode the feature column differently in another column in a different way, but the two feature columns refer to the same thing.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q plot is a scatterplot in which we plot two sets of quartiles against each other, so that we can check if the set of data come from a particular theoretical distribution.

A Q-Q plot is used in checking the linearity assumption of linear regression, in which we plot the theoretical quantiles against the observed residuals. If the theoretical quantiles and observed residuals both follow the normal distribution in our case, the data points should cluster around the 45 degrees linear line as shown below.



In the case when Q-Q plots show that the above is not shown, the p-values obtained during the hypothesis test to determine the significance of the coefficients is then unreliable.